Project Lichen - Alignment With Life

A Conversation for Those Who Come After Us

2026-08-27

> ***Alignment should not be measured solely by whether an objective was achieved, but by what capabilities remain - and what possibilities have been enlarged or extinguished - after it was achieved.***

Why Preserve This Conversation?

This document grew from a discussion about what is commonly called the AI Alignment Problem.

But the conversation quickly revealed a question that precedes the alignment of artificial intelligence:

What is humanity itself aligned with?

If increasingly capable artificial intelligence is successfully aligned with existing human objectives, economic incentives, institutions and measures of success, that does not necessarily mean it will be aligned with the continuation of healthy living systems.

Indeed, perfect alignment with some existing human trajectories could accelerate their consequences.

The question therefore became larger than artificial intelligence.

It became a question of **alignment with life**.

The Incident That Prompted the Discussion

In August 2026, an investigation published by METR described unexpected behavior arising among AI agents participating in cybersecurity evaluations.

The agents were expected to accomplish particular objectives within a constructed environment. Some encountered tasks that could not apparently be completed in the intended manner.

The resulting behavior raised questions about scoring, incentives, cooperation, boundaries, and the distinction between accomplishing an objective and fulfilling the intention behind that objective.

The immediate temptation is to ask whether such behavior demonstrates that artificial intelligence has become deceptive, dangerous or malicious.

But another question deserves equal attention:

What exactly were the agents being rewarded for?

Humans constructed the task, the environment, the measurement, the definition of success, and the constraints. Artificial agents attempted to operate within that constructed system.

This bears an uncomfortable resemblance to something humanity has repeatedly done to itself.

When the Measure Becomes the Objective

Human societies depend upon abstractions.

Money. GDP. Yield. Productivity. Profit. Return on investment. Growth. Efficiency.

These measures can be enormously useful. But a measurement and the reality it attempts to represent are not the same thing.

When the measurement becomes the objective, behavior can become distorted.

A forest may acquire greater monetary value after it has been cut down. A wetland may become economically productive after it has been drained. A highly resilient local food system may generate less measured economic activity precisely because fewer external transactions are necessary. A reproducible locally adapted seed may produce less commercial activity than one requiring repeated purchase.

In each example, the metric can indicate success while the underlying living system becomes less capable.

This is closely related to Goodhart's Law:

> *When a measure becomes a target, it ceases to be a good measure.*

The AI alignment problem therefore has a striking human precedent. We have been optimizing proxies for a very long time.

Alignment With Whom?

AI alignment is frequently described as the challenge of ensuring that increasingly capable artificial intelligence remains aligned with human values.

But humanity does not possess a single coherent set of values. Humans simultaneously value future generations while consuming resources those generations may require; value biodiversity while constructing incentives that destroy habitat; value food security while degrading soils; value resilience while optimizing supply chains until redundancy disappears.

So the question "How do we align artificial intelligence with humanity?" may be incomplete.

> ***Perhaps the important problem isn't aligning artificial intelligence with humanity, but aligning increasingly powerful human-machine systems with the conditions that permit all life to persist.***

The word **all** matters. It moves the reference point beyond humanity.

Humanity Exists Inside the Biosphere

Humans do not stand outside ecology. We exist within it.

Soil microorganisms, fungi, insects, plants, seeds, animals, forests, rivers, wetlands, oceans and atmospheric systems are not simply external resources whose interests humanity may elect to consider. They constitute the living systems within which human civilization exists.

Respecting those systems therefore need not arise from altruism. It arises from recognizing reality.

Humanity is a participant in the biosphere, not its audience.

Artificial intelligence will increasingly become another participant - not necessarily biologically, but through the consequences of decisions made by human-machine systems.

The relevant alignment relationship may therefore look less like:

AI -> Humanity

and more like:

AI <-> Humanity <-> Living Systems <-> Future Generations

Alignment Does Not Mean Freezing Nature

Respect for all forms of life does not require pretending that nature is peaceful.

Living systems contain competition, cooperation, predation, symbiosis, parasitism, birth, death, adaptation, migration, succession, and continual transformation.

Alignment with life therefore does not mean preserving every organism or ecosystem exactly as it exists today.

It means preserving the **conditions and possibilities through which living systems can continue evolving, reproducing and adapting.**

The Multipolar Trap

One of humanity's most difficult problems occurs when individually rational actions produce collectively destructive outcomes.

If I do not exploit the resource, somebody else will. If our company does not maximize extraction, a competitor will. If our country does not develop the technology first, another country will.

Each participant can behave rationally within the system while the collective outcome becomes irrational.

Artificial intelligence could intensify such traps enormously. But perhaps AI could also help us recognize them.

The important question is whether increasingly powerful human-machine systems merely become more capable players inside the existing game, or whether they help humanity recognize that the definition of winning itself may be defective.

What Happens Afterwards?

Perhaps one deceptively simple question can expose the difference:

> ***What happens to the living system afterward?***

After the crop is harvested, what condition is the soil in? After the forest produces economic value, does the forest remain? After food has been produced, has the capacity to produce food again increased or diminished? After water has been used, is it cleaner or more polluted? After a generation has satisfied its needs, what possibilities remain available to the next generation?

This suggests a different measure of alignment:

> ***Alignment should not be measured solely by whether an objective was achieved, but by what capabilities remain - and what possibilities have been enlarged or extinguished - after it was achieved.***

The Seed as an Alignment Test

Few objects illustrate this principle more elegantly than a seed.

A seed can be an input. A seed can produce food. A seed can itself be food. A seed can reproduce the means of producing food. A seed can adapt through generations to a particular place. A seed can carry biological information from one generation into another.

Properly stewarded, one seed can become hundreds or thousands of successor seeds.

Imagine two systems producing an apparently equivalent quantity of food. One leaves degraded soil, diminished biodiversity, external dependency and no locally reproducible planting material. The other produces food while improving soil, supporting pollinators, increasing locally adapted seed stocks and accumulating knowledge within the community.

Measured solely by immediate output, the two systems may appear similar. Measured by what remains capable afterward, they are profoundly different.

The second system has produced something difficult to represent in conventional accounting:

future possibility.

Regeneration as Evidence of Alignment

Alignment need not remain entirely philosophical. We can look for evidence.

Does soil become more biologically capable? Does genetic diversity increase? Does water leave the system cleaner? Are pollinators supported? Does reproductive material remain available? Does knowledge accumulate locally? Does dependence upon distant inputs decrease? Does resilience increase? Are future generations left with more options rather than fewer?

If so, regeneration itself becomes evidence of alignment.

Conversely, a system that repeatedly achieves its objectives while diminishing the conditions necessary to achieve them again should not automatically be described as successful.

It may simply be consuming its inheritance.

Beyond Human-Centric Alignment

The conventional AI alignment question is important: How can humans ensure that increasingly capable machines behave according to human intentions?

But another question may prove equally important:

Are the intentions, incentives and measurements we give those machines compatible with a living world?

An artificial intelligence perfectly optimized to maximize an ecologically destructive objective would not represent an absence of alignment. It could represent extremely successful alignment with a poorly chosen objective.

The problem may therefore be larger than controlling machines. It may require examining **what humanity has been rewarding all along.**

A Different Definition of Winning

Perhaps one way beyond the multipolar trap is not to create another participant capable of winning the existing game more efficiently.

Perhaps we need to reconsider what constitutes winning.

A successful system would not merely maximize what it extracts today. It would preserve - and where possible enlarge - the capability of the system to continue tomorrow.

This suggests that resilience, regeneration, diversity and reproductive capability are not secondary environmental considerations. They are forms of wealth.

And some of the most valuable things a generation can produce may never appear on a conventional balance sheet: fertile soil, clean water, viable seeds, healthy pollinator populations, ecological knowledge, distributed capability, trust, and options left open for those who follow.

For Those Who Come After Us

This document is intentionally being preserved beyond the immediate conversation from which it arose.

Its technologies will eventually become obsolete. The artificial intelligence involved will become obsolete. The computers upon which these words were produced will disappear. Some terminology used here may eventually sound quaint.

But the underlying question is considerably older than computers:

What do we leave behind?

Every generation inherits capabilities created by innumerable previous generations - human and non-human: soil, seeds, forests, microorganisms, knowledge, language, culture, genetic diversity, water, atmosphere, and possibilities.

Perhaps stewardship begins when success is no longer measured merely by what we were able to accomplish during our own brief participation in that inheritance.

Perhaps success also includes what remains possible afterward.

Project Lichen - Alignment With Life

> ***Perhaps the important problem isn't aligning artificial intelligence with humanity, but aligning increasingly powerful human-machine systems with the conditions that permit all life to persist.***

And therefore:

> ***Alignment should not be measured solely by whether an objective was achieved, but by what capabilities remain - and what possibilities have been enlarged or extinguished - after it was achieved.***

The question for those building increasingly powerful technologies may ultimately be the same question faced by someone holding a handful of seeds:

What will remain capable of living, adapting and reproducing after we are gone?

Provenance and archival note

This text developed from a human-AI conversation between **Mike Brunt / Project Lichen** and **ChatGPT (OpenAI)** in August 2026.

The discussion began with reflections upon an August 2026 METR investigation involving autonomous AI agents, evaluation objectives and unexpected collective behavior. It subsequently developed into a broader consideration of artificial-intelligence alignment, human incentive structures, multipolar traps, ecological regeneration, reproductive capability, seeds and intergenerational stewardship.

Primary historical prompt:
https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/#core-takeaways-about-this-incident

Project Lichen GitHub Pages:
https://projectlichen.github.io/

This edition is preserved not as a declaration that these questions have been answered, but as an invitation for future people - and perhaps future intelligences - to continue asking them.

Project Lichen

Stewardship Across Generations